OFROM
Corpus oral de français de Suisse romande



Mathieu Avanzi[1,2], Marie-José Béguelin[1], Federica Diémoz[1]
[1]Université de Neuchâtel ; [2]University of Cambridge

Version 2.2
Août 2015

Avertissement

Avanzi, M., Béguelin, M.-J. Diémoz, F. (2012-2015). « OFROM – corpus oral de français de Suisse romande, v. 2.2 », Ms, Université de Neuchâtel, http://www.unine.ch/ofrom

Avanzi, M., Béguelin, M.-J. Diémoz, F. (2015, à par.). « De l'archive de parole au corpus de référence : La base de données orale du français de Suisse romande (OFROM) », *Cahiers Corpus*.

Table des matières

1. Avant-propos

OFROM constitue la première archive comprenant uniquement des enregistrements de français parlé en Suisse romande aligné texte/son[1]. Les enregistrements que la base contient sont pour la plupart extraits d'entretiens guidés, à dominante monologique, dans lesquels l'interviewé (un locuteur né en Suisse, et vivant en Suisse romande) était sollicité pour répondre à des questions nécessitant des réponses plus ou moins longues posées par l'interviewer (le responsable de l'enquête). Une plus petite partie des enregistrements ressemblent davantage à des interactions, puisqu'ils impliquent au moins deux personnes qui parlent à bâtons rompus. Les thèmes abordés concernent aussi bien les métiers, les voyages, les passe-temps des locuteurs, que leurs relations de voisinage, leurs projets ou les situations incongrues auxquelles ils ont été confrontés dans leur vie. Elles peuvent également être en rapport avec le système politique ou la situation linguistique de la Suisse, etc.

Les enregistrements actuellement mis à disposition ont été réalisés à partir de 2008 par des étudiants de Bachelor, lors d'un travail conduit dans le cadre des séminaires de linguistique française dispensés pour les uns par Mathieu Avanzi et Marie-José Béguelin à l'Université de Neuchâtel (cote UNINE) ; pour les autres par Alain Berrendonner à l'Université de Fribourg (cote UNIFR). Durant cette période, les transcriptions associées aux fichiers sons ont été faites par les étudiants responsables de l'enquête. Elles ont toutes été vérifiées, anonymisées et uniformisées par un étudiant de Master avant leur mise en ligne. Depuis 2014, les enregistrements sont réalisés et transcrits par des collaborateurs scientifiques du Centre de dialectologie et d'étude du français régional de l'Université de Neuchâtel, dirigé par Federica Diémoz. En moyenne, les entretiens enregistrés durent entre 30 et 40 minutes, mais seules une dizaine de minutes sont transcrites pour chacun des locuteurs de la base. Des erreurs dans les transcriptions n'étant pas impossibles, nous invitons les utilisateurs de la base à nous signaler les éventuelles erreurs qui pourraient demeurer.

En mettant au point cette base, nous n'avons pas cherché à construire un « corpus de référence » du français parlé en Suisse romande, qui serait échantillonné en vertu de critères sociolinguistiques classiques comme le genre de discours, l'âge, le sexe et l'origine des locuteurs. L'entreprise aurait été trop difficile. Avec OFROM, nous avons simplement souhaité mettre à disposition de la communauté une base de données comprenant des enregistrements et les fichiers de transcription correspondants, sur laquelle il est possible de procéder

[1] La base de données Phonologie du Français Contemporain (PFC, *cf.* Durand et al. [2002 ; 2009]) héberge également des enregistrements de locuteurs romands, originaires de Nyon, de Genève et de Neuchâtel.

à des requêtes simples à l'aide d'un concordancier mis en ligne sur un site web convivial. Nous souhaitons que la base de données OFROM puisse servir à la recherche sur le français tel qu'il est parlé en Suisse romande.

2. Transcription des enregistrements

2.1. Support de transcription

À l'origine, les enregistrements ont été réalisés en vue d'études sur la prosodie, qui impliquent un alignement fin en phonèmes et syllabes, puis un codage des proéminences et des groupes accentuels de différents rangs [Avanzi 2013]. Dans ce contexte, nous avions décidé que les enregistrements seraient transcrits directement dans le logiciel Praat [Boersma & Weeninck 2012][2].

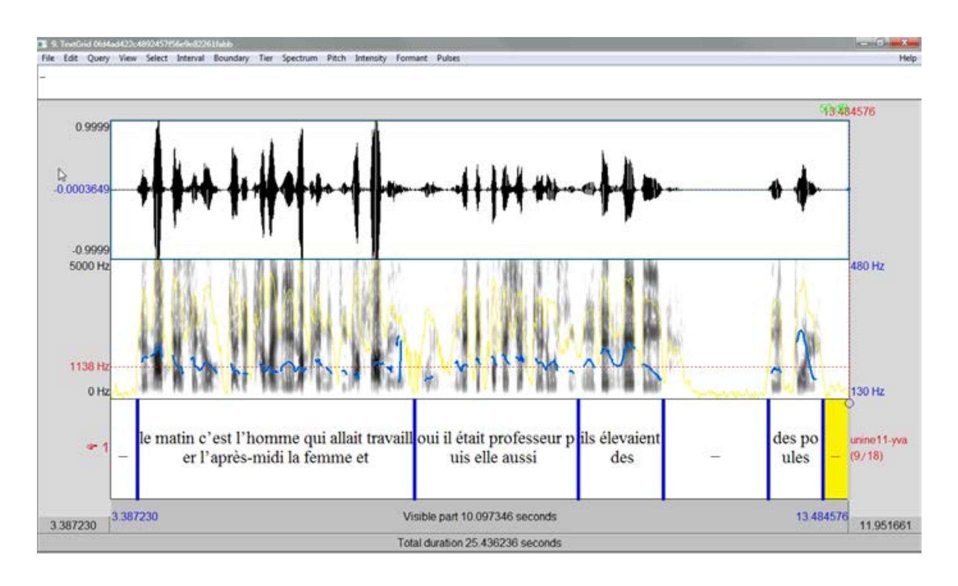


Figure 1. Copie d'écran Praat, avec dans la partie supérieure l'oscillogramme, dans la partie intermédiaire la courbe de F0, d'intensité et le spectre, et dans la partie inférieure, la transcription orthographique synchronisée sur le signal. Les pauses sont notées « _ » (*cf. infra* §2.3.5).

[2] Le logiciel Praat est gratuitement téléchargeable sur le web, et de nombreux tutoriels pour sa prise en main existent (cf. p. ex. http://latlcui.unige.ch/phonetique/praattutos.php pour un tutoriel en français). De par son ergonomie, son caractère gratuit et « open source », le logiciel Praat est aujourd'hui en passe de devenir le logiciel de référence pour les études de phonétique.

Signalons par ailleurs, à toutes fins utiles, qu'il existe de nombreux logiciels qui permettent de lire les fichiers de transcription au format TextGrid que génère Praat. Une fois convertis en xml (avec un logiciel comme Transformer [Ehmer 2006], les TextGrids peuvent être lus par des logiciels comme Transcriber [Barras et al. 1998] ou Clan [MacWhinney 2000].

Comme on peut le voir sur la Figure 1, le logiciel Praat permet en effet d'avoir accès à la transcription de façon alignée directement avec le son, et de visualiser de façon dynamique le spectre, la ligne d'intensité, les variations de F0 et les pauses de façon précise au cours du temps.

2.2. Identification des locuteurs

Les paroles d'un locuteur sont reproduites orthographiquement dans une « tire », c'est-à-dire dans une « couche de transcription » qui lui est propre, et qui porte son nom. Les locuteurs reçoivent chacun un nom de code, qui est unique dans la base de données. Ce code est composé de cinq lettres, suivies de deux chiffres, d'un trait d'union et de trois lettres. Les cinq premières lettres indiquent l'université dans laquelle l'étudiant était inscrit quand il a réalisé son enregistrement et sa transcription (UNINE pour Neuchâtel, UNIFR pour Fribourg), les chiffres qui suivent l'année universitaire pendant laquelle l'enregistrement a été réalisé (08 pour 2008, 09 pour 2009, etc.). Un trait d'union sépare ce premier code de trois autres lettres, dont les deux premières sont mises pour les initiales du locuteur ou du transcripteur (première lettre du prénom, même si c'est un prénom composé), la dernière pour différencier les locuteurs enregistrés une même année (« a » pour le premier locuteur, « b » pour le second, etc.). Ainsi, la locutrice dont la Figure 1 ci-dessus donne un extrait, a été enregistrée par un étudiant de l'université de Neuchâtel en 2011, a un prénom qui commence par « y » et un nom par « v », elle est la première locutrice ayant de telles initiales dans la volée d'enregistrements réalisés en 2011. D'autres informations ont été codées pour chacun des locuteurs de notre base de données. Elles peuvent être affichées dans des fenêtres spécifiques lors de la recherche sur le concordancier (*cf. infra*, §3).

À partir de 2014 le codage des locuteurs a été légèrement modifié. Les trois dernières lettres ont été remplacées par une simple numérotation sur trois chiffres allant de 001 à 999.

2.3. Conventions de transcription

2.3.1. Unités de transcription

Les empans ou intervalles de transcription dans les TextGrids Praat ne correspondent pas à des séquences linguistiques dont on serait en mesure de donner une définition scientifique stable. Les aligneurs automatiques nécessitant des empans de transcription relativement restreints [Goldman 2011 ; Bigi et al. 2012], nous avons pris le parti de sélectionner pour la transcription des fenêtres temporelles courtes, de 1 à 5 secondes au maximum. Au plan linguistique, ces séquences sont relativement sous-spécifiées : elles

correspondent à des groupes assortis d'une frontière intonosyntaxique mineure ou majeure (l'idée est que l'on ne coupe pas au milieu d'un mot ou d'un syntagme de bas rang).

2.3.2. Choix d'une orthographe standard

Les enregistrements sont transcrits en orthographe standard, sans « trucages » ni ponctuation : nos conventions suivent en cela les recommandations du GARS [Blanche-Benveniste & Jeanjean 1986 ; Blanche-Benveniste 1997], reprises dans la plupart des corpus de français parlé transcrits existants [DELIC 2004 ; Dister & Simon 2008 ; Branca et al. 2012]. Ces trucages orthographiques, qui sont largement illustrés dans les publications de référence [Giovanni & Savelli 1990], sont d'autant moins nécessaires que les aligneurs contiennent des dictionnaires qui incluent plusieurs variantes de prononciation pour un même mot [Adda-Decker & Snoeren 2011]. Cela dit, pour que les aligneurs fournissent des résultats optimaux, il faut que la transcription du texte « colle » au plus près à ce qui est prononcé. Par conséquent, nous avons dû prendre quelques distances quant aux autres conventions de transcription généralement suivies par nos collègues.

2.3.2.1. Règles typographiques

Compte tenu de la non-pertinence de la ponctuation orthographique pour transcrire l'oral, l'usage des majuscules est réservé aux noms propres, aux titres d'ouvrages (1), aux acronymes (2)-(3)[3] et aux lettres prononcées de façon isolées (4)-(5) :

(1) vous avez gagné un voyage à **Abidjan** _ et puis _ c'est la seule fois où je suis allée en avion nous sommes allés en **Côte d'Ivoire** _ et nous avons vécu des heures extraordinaires avec l'orchestre de **Lille** _ dirigé par **Casadesus** _ ils ont et puis ils ont joué _ l'air de **Nabucco** [unine11-yva]
(2) j'ai un **CFC** de commerce [unine08-rza]
(3) j'habitais près d'une stat/ euh une base aérienne de l'**OTAN** tenu par des américains [unine11-rpa]
(4) un nom qui commence avec un **W** [unine11-sma]
(5) tous les films qu'ils voient c'est des films euh de série **B** avec toujours des femmes blanches euh prêtes à coucher avec n'importe qui [unine09-lba]

Les mots prononcés de façon abrégée sont transcrits comme tels, sans apostrophe finale :

(6) et le **prof** était pas mal en plus [unine08-ema]

[3] La signification des acronymes n'est pas donnée dans la transcription.

(7) l'**uni** en sport prend pas mal de temps [unine08-oca]
(8) j'ai été à l'école de **com** [unine11-rpa]

Les mots prononcés de façon non abrégée sont transcrits dans leur forme pleine :

(9) voilà donc ça c'est pour euh les soirées **et cetera** on peut encore parler des journées [unine08-aha]

Les chiffres sont transcrits en toutes lettres, y compris les âges, les dates, les quantités, etc. :

(10) ben j'ai commencé à prendre des cours de chant quand j'avais **quinze** ans [unine11-vpa]
(11) et puis euh il est de **dix-neuf cent vingt-cinq** [unine11-eja]
(12) tu me la vendrais pas **quatre mille** francs [unine11-tpa]

Les mots étrangers sont transcrits dans leur orthographe d'origine. S'ils ne sont pas connus, ils ne sont pas transcrits (*cf. infra*, §2.3.4) :

(13) après une petite semaine euh de galère quand même euh en **couch surfing** [unine08-ema]
(14) on a passé euh ben avec mon copain on a passé le le **dive** euh **l'open water** en fait [unine11-nfa]
(15) mais après y a eu l'évolution dans l'entreprise où on nous a fait des **ordersatz** [unine11-jea]

2.3.2.2. Morphologie

La consigne donnée aux transcripteurs est que l'on transcrit ce que l'on entend, et que l'on ne transcrit pas ce que l'on n'entend pas. Ainsi, nous ne notons pas systématiquement tous les pronoms dans les tournures impersonnelles, *cf.* (16), ni les « ne » de négation si ceux-ci ne sont pas clairement audibles, *cf.* (17) et (18), respectivement :

(16) comme **il y a** un temps pour tout même en politique _ **y a** un temps pour tout donc moi à Berne je vois plutôt des gens [unine11-eza]
(17) et avec ce jeune brass band qui **n'**a que quatre ans [unine11-gpa]
(18) euh des des gens qui sont **pas** forcément expérimentés là-dedans [unine08-sea]

Nous retranscrivons l'élision quand elle est réalisée, nous ne la transcrivons pas lorsqu'elle ne l'est pas :

(19) elle s'éloigne un peu de la _ de la voiture et pis lui en mettant le contact en s'approchant de _ **d'elle** et ben la voiture elle explose [unine11-jva]
(20) ou même ce qui se passe autour **de elle** [unine11-sca]

Nous ne changeons pas non plus la forme morphologique du mot si une règle d'accord en genre ou en nombre n'est pas respectée (21)-(22), ou si le pronom ou le mode du verbe enfreint la norme, *cf.* respectivement (23)-(24) :

(21) là euh je sais pas la Sagrada Familia c'est une énorme église que Gaudi **a fait** [unine11-tpa]
(22) ça devient une euh **une** mécanisme de groupe [unine11-sma]
(23) je vais beaucoup moins parce que j'**y** dis faut aussi un peu te calmer maintenant [unine11-jsa]
(24) et puis euh chaque aide-soignante **s'asseye** à côté d'une personne pour l'aider à prendre son repas [unine09-tba]

Enfin, nous ne signalons pas les écarts de prononciation dans la transcription, que ces écarts soient courants ou non. Par exemple, « parce que » est toujours écrit tel quel, qu'il soit prononcé [πασκ↔] ou [πα®σκ], de même que des morphèmes comme « enfin » prononcé [φE®].

2.3.3. Amorces

Les amorces de mots sont signalées par des slash « / » qui suivent les premières lettres du morphème inachevé :

(25) y a d'excellentes **boulan/** boucheries aussi tout au long du vallon [unine11-rpa]
(26) y a un grand juré euh qui **habit/** qui qui au **b/** aux Etats-Unis ils ont euh [unine11-jva]

2.3.4. Segments non transcrits

Nous codons « % » certaines portions de signal que nous ne transcrivons pas. Certains éléments ne sont pas transcrits car le ou les mots prononcés sont incompréhensibles (en raison d'une mauvaise articulation, d'un changement de qualité vocale, d'un chevauchement de parole) ou à des fins d'anonymisation (*cf. infra*, §2.4).

2.3.5. Pauses vides et pauses « pleines »

Les informations relatives à l'habillage suprasegmental sont directement lisibles dans Praat. Nous ne les indiquons donc pas dans nos transcriptions. Nous notons cependant de façon systématique les pauses silencieuses, et ce peu

importe leur durée. Les pauses silencieuses sont ainsi isolées dans des intervalles dédiées et transcrites à l'aide du symbole « _ »[4]. Nous avons été moins précis pour les pauses remplies (allongements et *euh* associés à des hésitations), qui ne sont pas forcément cantonnées dans des intervalles dédiés, mais comprises dans les mêmes intervalles que les mots auxquels elles s'accolent.

2.4. Anonymisation

La parole est une propriété [Baude 2006]. Les locuteurs enregistrés dans notre corpus ont signé des autorisations stipulant qu'ils donnaient leur accord pour l'enregistrement, la diffusion et l'analyse, à des fins linguistiques, de leur parole, à condition que les données soient anonymisées. Nous n'avons pas procédé à une anonymisation du signal à proprement parler. Pour éviter de rendre publiques certaines informations prononcées pouvant servir à l'identification des locuteurs, nous avons simplement fait correspondre aux séquences sonores pouvant aider à l'identification du locuteur des intervalles dédiés à l'intérieur de la couche de transcription. Ces intervalles contiennent un symbole spécial (« # »), qui permet, lors de la recherche à l'aide du concordancier, que le contenu sonore associé à l'intervalle incriminé ne puisse pas être entendu ni téléchargé (*cf. infra*, §4). L'anonymat des locuteurs de notre corpus est ainsi préservé.

3. Données métalinguistiques

Les enquêteurs avaient pour consigne de recueillir, pour chaque locuteur enregistré, un certain nombre d'informations qui devaient permettre de trier les locuteurs selon des critères sociolinguistiques minimaux au moment de la recherche sur concordancier. Ces critères seront détaillés plus loin (*cf.* §4.2.2).

4. Mode d'emploi du concordancier

4.1. Avertissement

Pour des raisons d'anonymisation, il n'est pas possible de télécharger l'ensemble des enregistrements et des transcriptions de la base, ni même un fichier sonore et la transcription complète associés à un locuteur ou un groupe de locuteurs. Pour avoir accès au contenu de la base, il est donc obligatoire de passer par le moteur de recherche.

---

[4] Leur durée est indiquée dans des info-bulles lors de la recherche sur le concordancier (*cf. infra* §4.2.6.1.).

4.2.Concordancier

Le moteur de recherche se présente sous la forme d'un « concordancier », grâce auquel il est possible de chercher une suite de caractère, un item monolexical ou polylexical, une étiquette, un lemme ou une combinaison de ces informations, dans un contexte donné. Le concordancier d'OFROM se présente de la façon suivante :

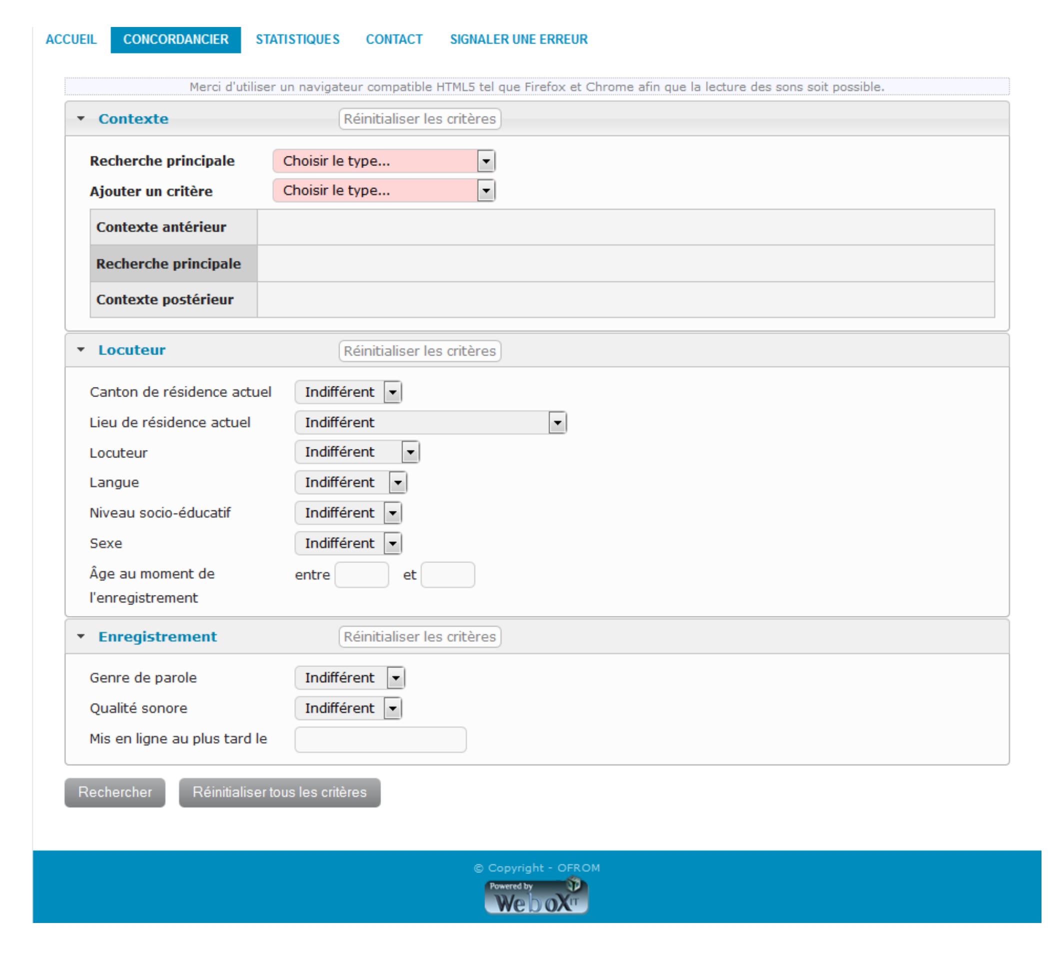


Figure 2. Aperçu d'ensemble du concordancier OFROM.

L'utilisateur doit d'abord choisir le type d'élément qu'il souhaite rechercher, en cliquant sur la première boite rose en haut de la fenêtre, « Choisir le type… ». S'il choisit de chercher un mot entier (token), une chaîne de caractère ou un lemme[5], une nouvelle case (en vert), apparaît sur la droite, il suffit alors

[5] Les lemmes correspondent à la forme non-fléchie des adjectifs, noms, déterminants, etc. et la forme à l'infinitif des verbes.

d'entrer le texte voulu et de cliquer sur le bouton « rechercher » en bas de la page :

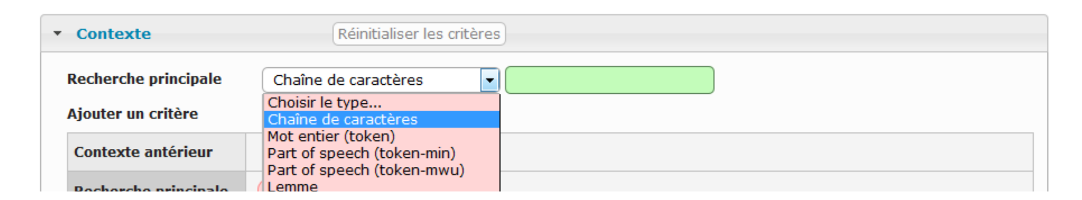


Figure 3. Fonction « recherche » du concordancier.

Si l'utilisateur choisit de faire une recherche par étiquette morphosyntaxique, alors trois nouvelles cases apparaissent. Dans les cases oranges, sur la droite, il est possible de sélectionner la catégorie d'étiquette (Adjectif, Pronom, Verbe, etc.), et dans la seconde de raffiner la recherche par sous-type d'étiquette (Verbe conditionnel auxiliaire, Pronom personnel sujet, etc.). S'il le désire, il peut également spécifier dans la case verte, la forme du texte recherchée (p.ex. on peut rechercher tous les éléments pronoms personnels objets qui ont la forme *le*).

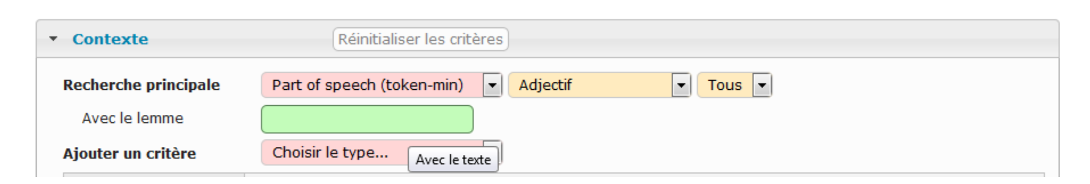


Figure 4. Fonction « recherche » du concordancier.

Comme on peut le voir sur la Figure 2, il est possible d'affiner la recherche. Pour ce faire, trois types de filtres peuvent être appliqués.

NB : il est possible de réinitialiser les critères de recherche pour chacun des trois filtres. On peut également réinitialiser tous les critères en une seule fois.

4.2.1. Filtre contexte

Avec le premier filtre, on peut restreindre le contexte antérieur et postérieur de l'unité en « ajoutant un critère ». Les mêmes options que celles proposées pour la recherche principale sont alors disponibles :

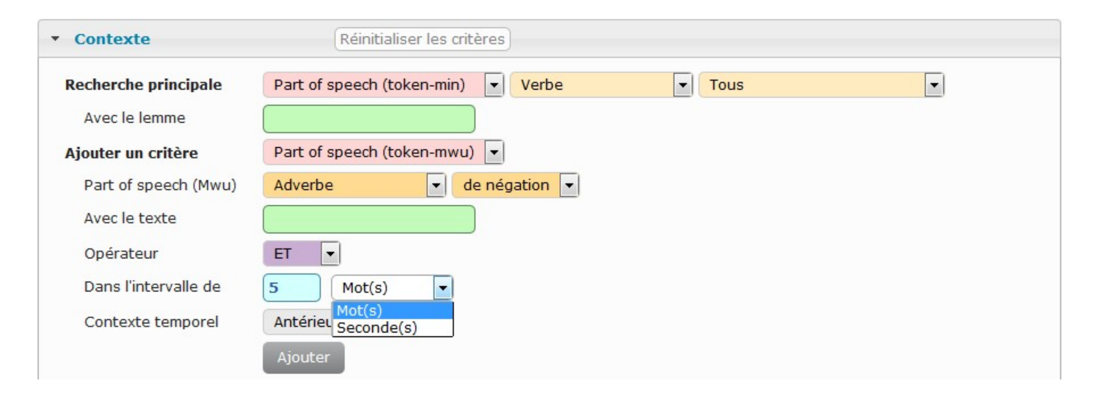


Figure 5. Fonction « recherche avec contexte » du concordancier.

On peut alors préciser l'empan temporel du contexte de recherche (en termes de nombre de mots ou de secondes), et préciser si on veut que l'élément soit avant ou après l'élément principal recherché. Il faut ensuite cliquer sur le bouton « ajouter », et alors on peut de nouveau préciser le contexte, en ajoutant d'autres filtres :

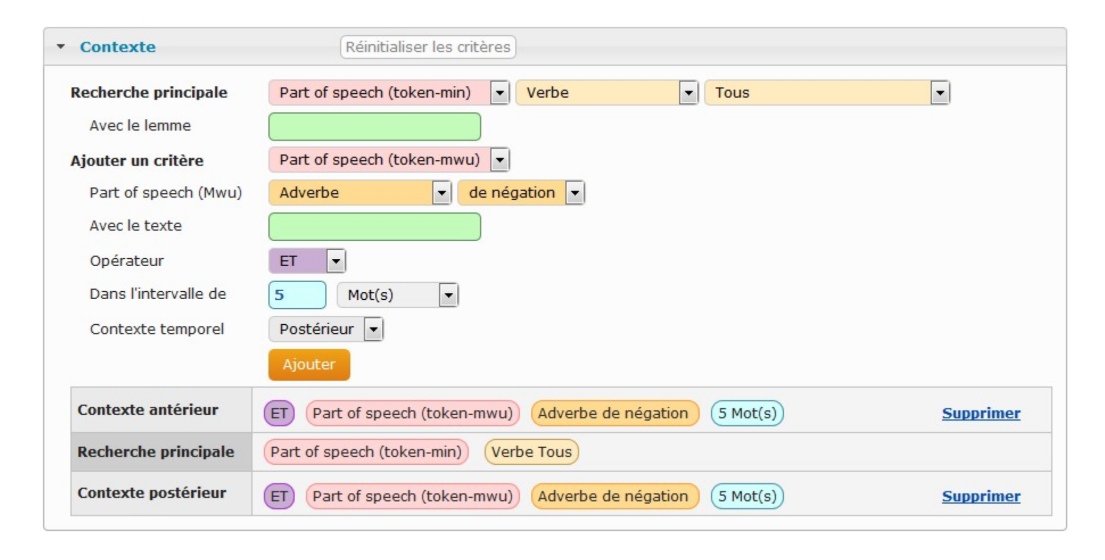


Figure 6. Fonctions « filtre contexte » du concordancier.

La syntaxe de la requête est alors visible dans des lignes en-dessous des filtres. Les opérateurs logiques (en violet) sont au nombre de trois : ET, OU ou SAUF. On peut supprimer une ligne de requête en cliquant sur le bouton « supprimer » de la ligne en question.

NB : Dans le cas où l'on cherche des chaînes de mots très précises, il faut bien préciser l'ordre des éléments du contexte par rapport au mot cible. Ainsi, si l'utilisateur travaille sur l'ordre des clitiques et qu'il cherche tous les pronoms personnels sujets suivis d'un « ne » de négation et d'un pronom personnel objet, il faudra qu'il spécifie intervalle +1 mot pour l'adverbe de négation, et l'intervalle + 2 mots pour le pronom personnel objet :

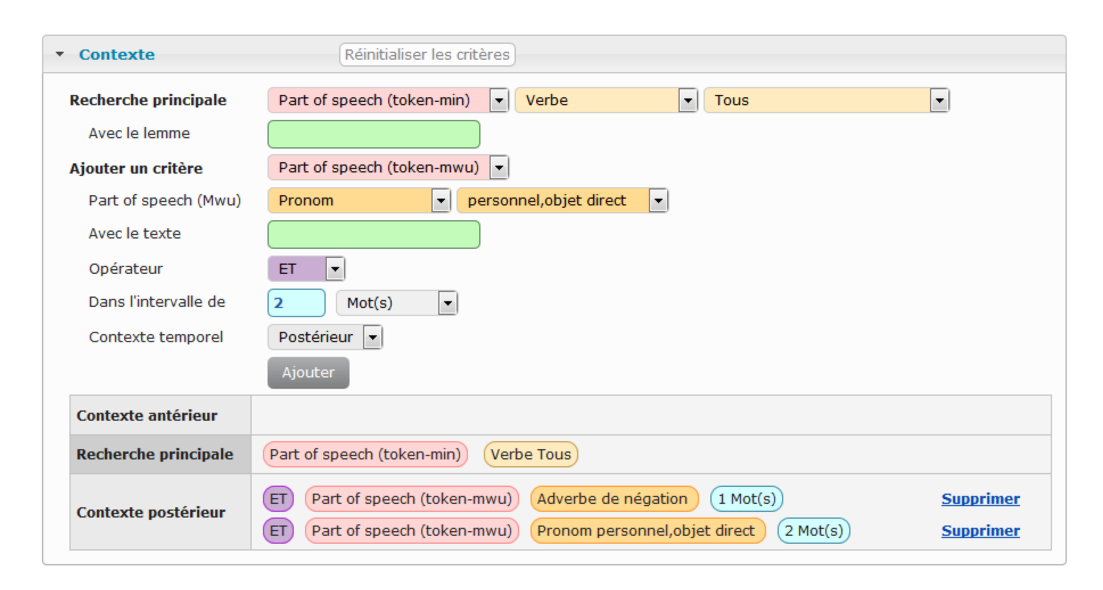


Figure 7. Exemple de recherche de la séquence [pronom personnel sujet + adverbe de négation + pronom personnel objet].

4.2.2. Filtre locuteur

On peut également définir une catégorie de locuteurs selon des critères sociolinguistiques classiques. Ces critères sont extrêmement rudimentaires et ne permettent pas de faire une recherche poussée en fonction de l'identité sociolinguistique des locuteurs. Ils permettent toutefois de procéder à un tri grossier des données. Chacune des rubriques contient un menu déroulant qui s'ajuste automatiquement en fonction des choix que l'on opère lors de l'application de tel ou tel critère.

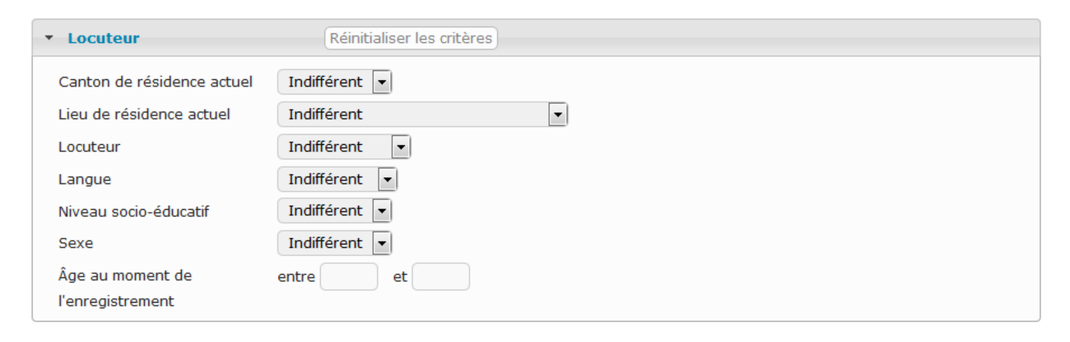


Figure 8. Fonctions « filtre locuteur » du concordancier.

Comme on le voit sur la Figure 8, on peut ainsi sélectionner les locuteurs selon leur canton[6] et/ou leur lieu de résidence actuel. On peut également sélectionner un ou plusieurs locuteurs de la base si on connait leur nom de code. On peut aussi filtrer les locuteurs selon qu'ils sont francophones natifs (L1) ou non (L2), selon leur niveau socio-éducatif[7], selon qu'il s'agit d'un homme ou d'une femme. On peut également indiquer un intervalle temporel pour spécifier la tranche dans laquelle le locuteur est né.

4.2.3. Filtre enregistrement

Enfin, il est possible de filtrer la recherche selon le genre de parole et la qualité sonore de l'enregistrement :



Figure 9. Fonctions « filtre enregistrement » du concordancier.

Pour le moment, la base de données OFROM ne contient que des interviews à dominante monologique et des dialogues, que nous avons catégorisés comme des « narrations » ou des « discussions ». On trouve également un extrait de conférence. Dans le futur, des enregistrements d'autres genres devraient être disponibles.

[6] Selon l'un des 7 cantons suisses où le français est langue officielle, soit Genève, Vaud, Berne, Neuchâtel, Valais, Jura et Fribourg.

[7] Les niveaux socio-éducatifs sont les suivants : niveau 1 : école obligatoire avec apprentissage plutôt technique ; niveau 2 : école obligatoire avec apprentissage plutôt bureau ; niveau 3 : maturité ; niveau 4 : études universitaires. Cette catégorisation est celle qui a été utilisée pour identifier socio-éducativement les locuteurs du point Neuchâtel qui figurent sur le site du projet PFC (*cf.* pour une présentation de ces données, Racine & Andreassen [2012]).

4.2.4. Visualisation des résultats de la recherche

Les résultats de la recherche s'affichent dans une nouvelle fenêtre, qui, lorsqu'elle apparaît, masque le concordancier :

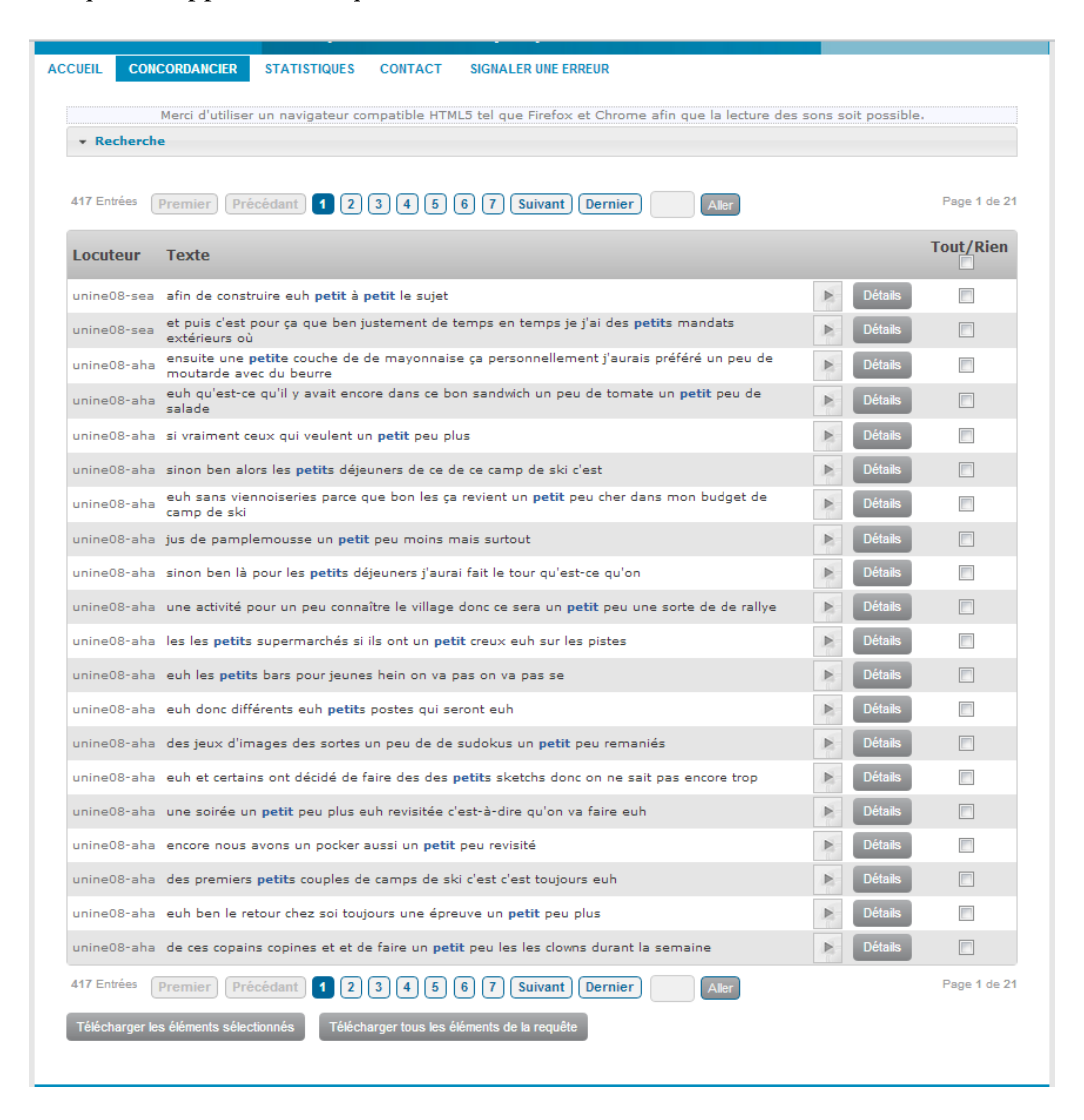
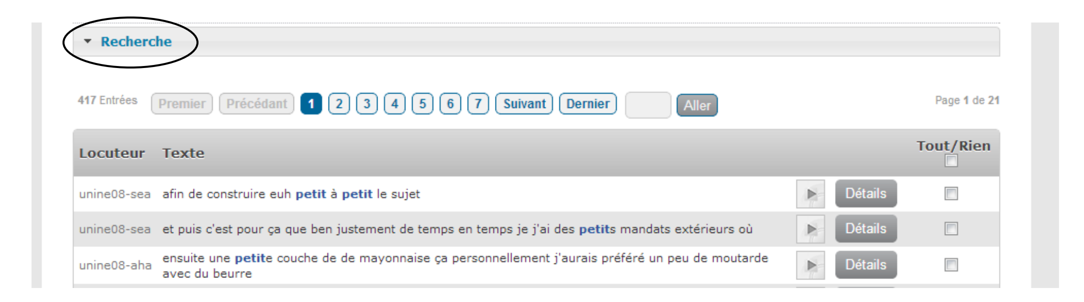


Figure 10. Résultats de la recherche de l'adjectif « petit » (première page).

L'élément recherché (ici l'adjectif *petit*) est mis en évidence (en gras et en bleu), et le nom de code du locuteur est donné sur la même ligne que l'extrait, juste devant le texte (en grisé).

NB : il est possible d'afficher de nouveau les critères de recherche, pour changer ou raffiner la requête en cliquant sur le bouton « recherche » en haut à gauche de l'écran.

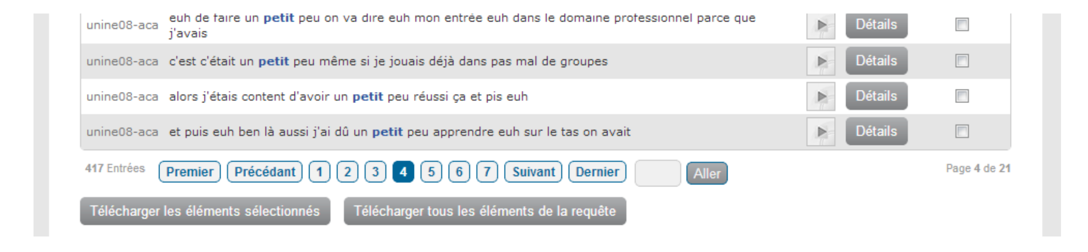


À partir de là, on peut écouter en cliquant sur le bouton lecture (▶) l'élément dans son contexte immédiat, tel qu'il est affiché à l'écran, ou bien cliquer sur « détails » pour accéder à plus d'options (*v. infra*, §4.2.6).

NB : pour que la lecture des sons soit possible, il est conseillé d'utiliser un navigateur compatible HTML5 (tel que Firefox ou Chrome).

4.2.5. Extraction des fichiers associés à la recherche

Il est possible de sélectionner certains éléments ou l'ensemble des éléments trouvés pour sa recherche en cochant les cases à la fin des lignes où apparaît chaque occurrence. Le nombre de résultats est limité à 20 par page. On peut naviguer dans les pages grâce à des commandes en dessous et au-dessus des résultats.



Une commande, au bas de la page permet de télécharger le ou les éléments cochés, ou tous les éléments de la requête.

Lorsque l'on procède au téléchargement de la sélection, on télécharge un fichier zippé qui contient plusieurs fichiers : un fichier au format wav et sa transcription au format TextGrid, qui correspondent à l'extrait affiché. Un fichier au format csv (que l'on peut ouvrir dans un tableur, type Excel) qui contient le nom de code du locuteur, les informations métalinguistiques

associées au locuteur et le contenu de la séquence (s'il y a plusieurs éléments sélectionnés, le tableur comprend autant de lignes que d'éléments sélectionnés).

4.2.6. Détails de l'extrait

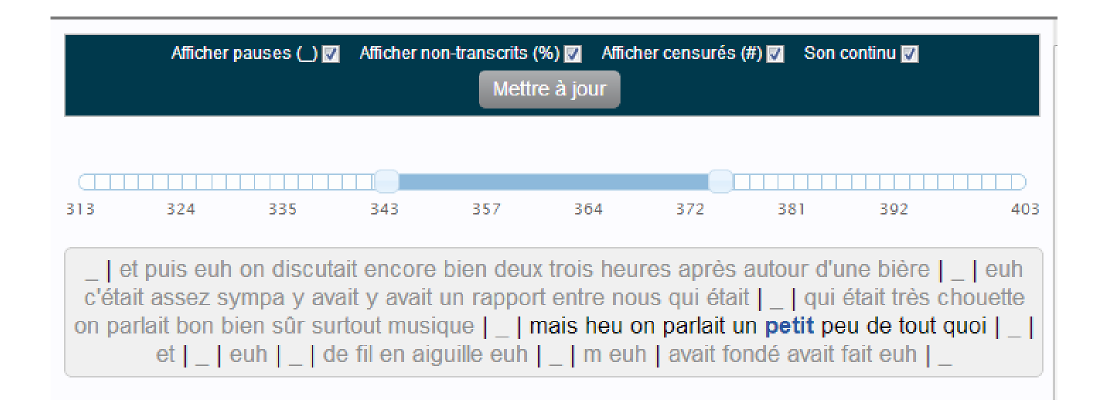
4.2.6.1. Agrandir/restreindre le contexte de l'élément recherché

Si l'on clique sur le bouton « détails » pour une occurrence donnée, une nouvelle fenêtre apparaît. Elle affiche la séquence dans son contexte, ainsi que d'autres options :



Figure 11. Fenêtre d'information de l'extrait (« détails »).

En haut de la fenêtre, une frise indique la position temporelle de l'extrait dans l'ensemble du fichier son, les chiffres affichent les secondes. On peut élargir ou restreindre le contexte de l'occurrence que l'on recherche. Pour que la modification ait lieu, il est impératif de « mettre à jour » en cliquant sur le bouton au-dessus de la frise :



NB : on peut choisir d'afficher ou non les pauses « _ », les segments non-transcrits (« % ») ou censurés (« # »), jouer le son en continu ou non.

NB : la durée des pauses s'affiche en millisecondes dans des info-bulles lorsque l'on passe la souris sur l'intervalle :



4.2.6.2. Lecteur intégré

Le lecteur intégré offre les fonctions classiques d'un lecteur de son : les touches ⏪ et ⏩ permettent de passer un intervalle ou d'y revenir, on peut lire (▶) ou arrêter (■) le son, changer le volume, lire en boucle, etc.

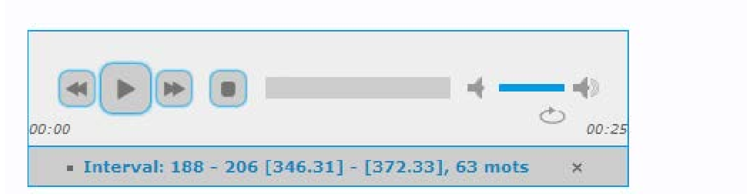
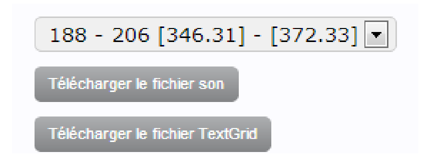


Figure 12. Lecteur de son intégré.

4.2.6.3. Téléchargement dynamique

Si l'utilisateur le désire, il est également possible d'entrer directement les bornes de début et de fin de l'extrait que l'on souhaite écouter :



Il est aussi possible de télécharger le fichier son au format wav et sa transcription au format TextGrid. Les fichiers sont générés dynamiquement, en fonction du contexte sélectionné : les extraits wav et TextGrids sont concaténés automatiquement en un seul fichier (y compris les pauses[8]), et peuvent être ensuite édités dans Praat.

NB : Ne pas oublier de mettre à jour la sélection avant de télécharger le son et ou le fichier TextGrid.

4.2.6.4. Informations sur le locuteur et sur l'enregistrement

Enfin, des informations sur le locuteur et l'enregistrement sont indiquées dans des tableaux en bas de la page.

Information sur le locuteur	
Nom	unine08-aca
Langue	Français L1
Sexe	Homme
Année de naissance	1953
Lieu de naissance	NR
Lieu de résidence actuel	NR
Canton de résidence actuel	NR
Y habite depuis	NR
Niveau socio-éducatif	Niveau 4
Occupation	Enseignant
Âge au moment de l'enregistrement	55
Statut familial	Marié

Informations sur l'enregistrement	
Qualité	Bonne
Type	m
Genre de parole	narration
Uni	UNINE
Propriétaire	Mathieu Avanzi
Lieu d'enregistrement	Fribourg
Canton	Fribourg
Enregistré par	Amélie Cochard
Date d'enregistrement	21/04/2008
Transcrit par	Amélie Cochard
Révisé par	François Delafontaine
Date de révision	18/09/2012

Figure 13. Informations sur le locuteur et sur l'enregistrement.

Les informations affichées étant assez transparentes, nous ne les commentons pas davantage ici.

[8] Les contenus sonores des intervalles codés « # » et « % » ne sont pas compilés dans le fichier final.

5. Statistiques

La rubrique « statistiques » permet de consulter, à tout moment, le contenu de la base de données.

Au moment de sa mise en ligne (décembre 2012), le corpus contenait 154’883 mots, était d’une durée d’environ 17 heures et compte 74 locuteurs, qui se répartissent de la façon suivante :

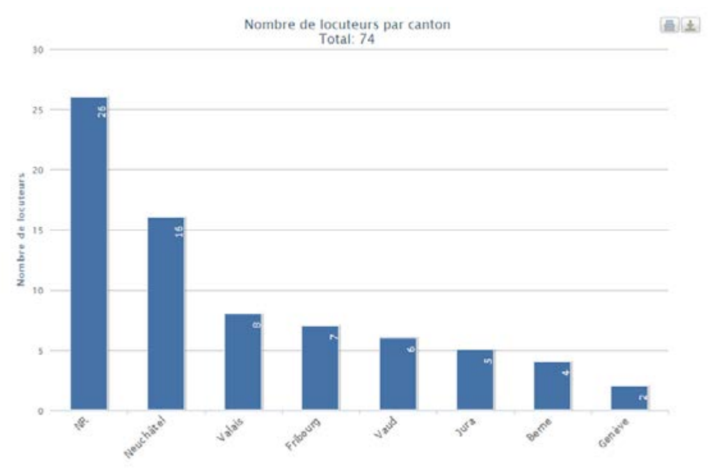


Figure 14. Nombre de locuteurs par canton dans la base de données au mois de décembre 2012.

En septembre 2013, la base s’enrichissait d’environ 50 nouveaux locuteurs. Le corpus était d’une durée de plus de 28 heures, et contenait environ 232’536 mots. Au total, la base contenait alors 119 locuteurs qui se répartissent de la façon suivante :

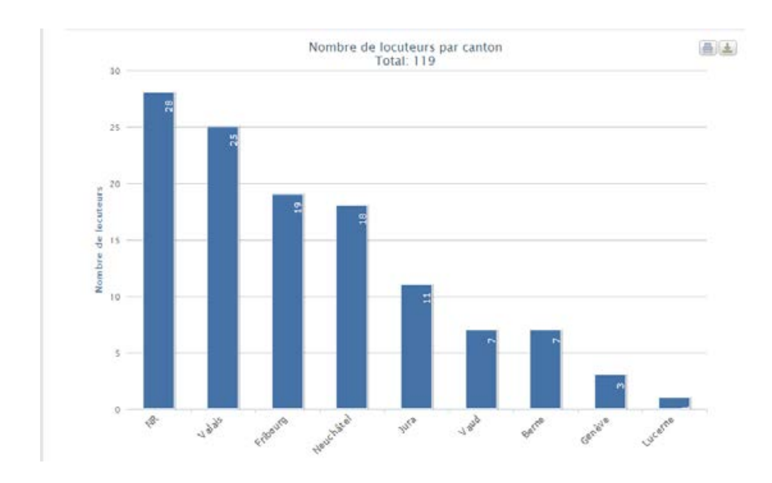


Figure 15. Nombre de locuteurs par canton dans la base de données au mois de sept. 2013.

Le corpus est, suite à la dernière mise à jour (août 2015), d'une durée d'un peu plus de 64 heures, et contient environ 615'621 mots pour un total de 222 locuteurs qui se répartissent de la façon suivante :

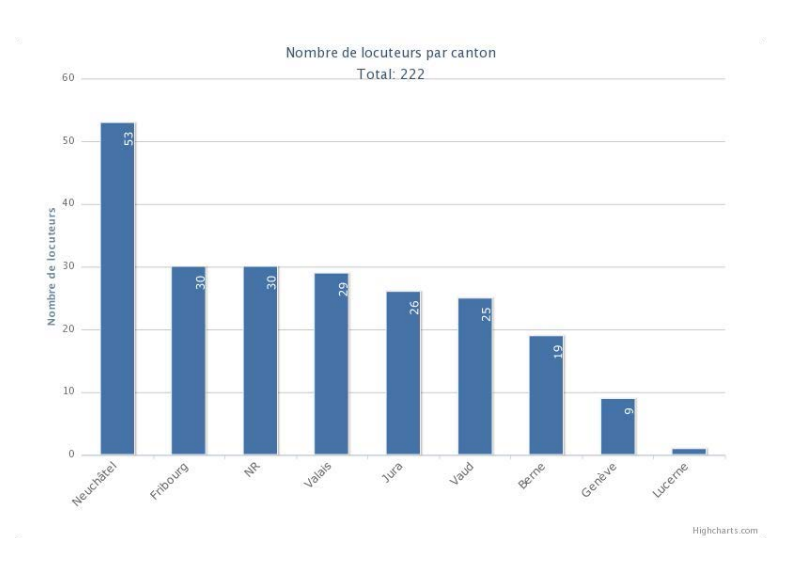


Figure 16. Nombre de locuteurs par canton dans la base de données au mois d'août 2015.

6. Formulaire de contact ou pour signaler une erreur

Un formulaire de contact est à disposition de l'utilisateur, de même qu'un formulaire permettant de signaler une erreur

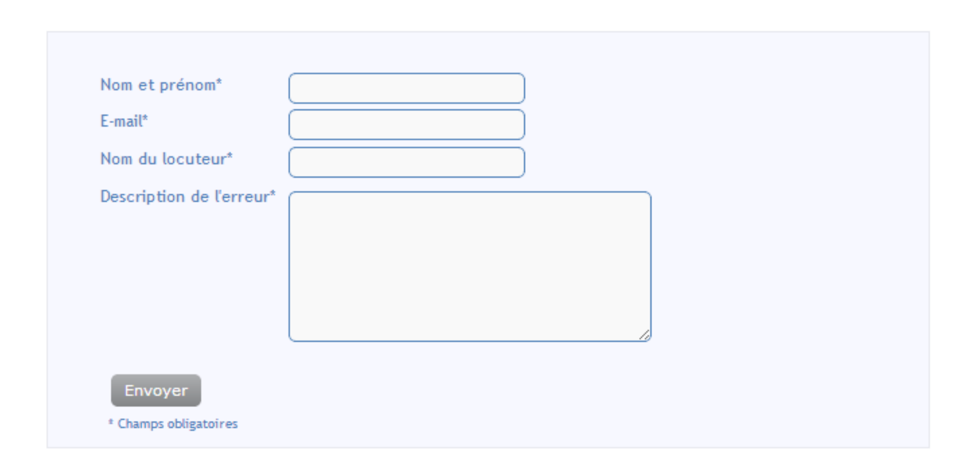


Figure 17. Formulaire de contact pour signaler une erreur.

7. Remerciements

La confection de ce site n'aurait jamais été possible sans le soutien financier du programme Campus virtuel Suisse, de la Faculté des Lettres et des Sciences Humaines de l'Université de Neuchâtel, ainsi que du Rectorat de l'Université de Neuchâtel et du Fonds National Suisse de la recherche scientifique (subside n°P300P1_147781).

Nous remercions Pierre Ménétrey (http://www.webox-it.com/), webmestre, pour le travail de confection du site. Merci également à Sandra Schwab (Université de Genève) pour ses conseils et la confection des scripts Praat qui ont été utilisés pour la mise en ligne des premières données sonores et des transcriptions associées. George Christodoulides nous a fourni le logiciel pour tagger la base de données et créer des fichiers xml pour la charger. François Delafontaine (Université de Neuchâtel) a réalisé un travail colossal de révision et de correction des transcriptions.

Nous remercions également toute l'équipe du Centre de dialectologie : Nathaniel Hiroz, Camille Legrand, Aline Widmer, Maude Ehinger et Julie Rothenbühler ainsi qu'Anna Schwab et Gwendoline Grivel pour les enquêtes de terrain et Julie Rothenbühler et François Delafontaine pour la transcription, la correction et la mise en ligne. Christophe Benzitoun a nettoyé, dans le cadre du projet ANR ORFEO, certains des fichiers présents dans la base.

Enfin nous remercions l'ensemble des collaborateurs scientifiques, des étudiants et des locuteurs qui ont participé aux diverses campagnes d'enquête.

8. Références

Adda-Decker, M. & Lamel, L. [1999]. "Pronunciation variants across system configuration, language and speaking style". *Speech Communication*, 29, 83-98.

Adda-Decker, M. & Snoeren, N. D. [2011]. "Quantifying temporal speech reduction in French using forced speech alignment", *Journal of Phonetics*, 39, 261-270.

Avanzi, M. [2013]. "Note de recherche sur l'accentuation et le phrasé prosodique à la lumière des corpus de français", *Tranel*, 5-24.

Barras, C. Geoffrois, E. Wu, Z. & Liberman, M. [1998]. "Transcriber: a Free Tool for Segmenting, Labeling and Transcribing Speech", *Proceedings of the First International Conference on Language Resources and Evaluation (LREC)*, 1373-1376.

Baude, O. (éd.). [2006]. *Corpus oraux. Guide des bonnes pratiques*. Paris: CNRS-Editions.

Bigi, B., Péri, P. & Bertrand, R. [2012]. "Influence de la transcription sur la phonétisation automatique de corpus oraux", *Actes des JEP*, 449-456

Blanche-Benveniste, Cl. & Jeanjean, C. [1986]. *Le français parlé. Edition et transcription*. Paris: Didier Erudition.

Blanche-Benveniste, Cl. [1997]. Approches de la langue parlée en français. Paris/Gap: Ophrys.

Boersma, P. & Weenink, D. [2012]. "Praat, v. 5.3". http://www.fon.hum.uva.nl/praat/.

Christodoulides, G., Avanzi, M., Goldman, J.-PH. [2014]. "DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech", Proc. LREC, 3902-3907.

Dister, A. & Simon, A. C. [2008]. "La transcription synchronisée des corpus oraux. Un aller-retour entre théorie, méthodologie et traitement informatisé", *Arena Romanistica* 1/1, 54-79.

Durand, J., Laks, B., Lyche, C. [2002]. "La phonologie du français contemporain : usages, variétés et structure". In *Romance Corpus Linguistics - Corpora and Spoken Language*, Pusch, C., Raible, W. (éds), Tübingen: Gunter Narr Verlag, 93-106.

Durand, J., Laks, B., Lyche, C. [2009]. *Phonologie, variation et accents du français*, Paris : Hermes.

Giovannoni, D. C. & Savelli, M.-J. [1990]. "Transcrire, traduire, orthographier le français parlé. De l'impossible copie à la falsification des données orales", *Recherches sur le français parlé*, 10, 19-37.

Goldman, J.-P. [2011]. "EasyAlign: an Automatic Phonetic Alignment Tool under Praat", *Proceedings of Interspeech*, 3233-3236.

MacWhinney, B. [2000]. *The CHILDES Project: Tools for Analyzing Talk*. 3rd Edition. Mahwah, NJ: Lawrence Erlbaum Associates.

Racine, I., & Andreassen, H. [2012]. "A phonological study of a Swiss French variety: Data from the canton of Neuchâtel", in R. Gess, C. Lyche, & T. Meisenburg (Eds), *Phonological Variation in French: Illustrations from Three Continents*, Amsterdam: John Benjamins, 173-207.